Current Biology 19, 1308–1313, August 11, 2009 ©2009 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2009.06.060

Report

Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain

Rodrigo Quian Quiroga,[1,2,3,*] Alexander Kraskov,[2,4] Christof Koch,[2] and Itzhak Fried[3,5]
[1]Department of Engineering, University of Leicester, LE1 7RH Leicester, UK
[2]Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125, USA
[3]Department of Neurosurgery, David Geffen School of Medicine, and Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA 90095-7039, USA
[4]UCL Institute of Neurology, WC1N 3BG London, UK
[5]Functional Neurosurgery Unit, Tel Aviv Medical Center and Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 64239, Israel

Summary

Different pictures of Marilyn Monroe can evoke the same percept, even if greatly modified as in Andy Warhol's famous portraits. But how does the brain recognize highly variable pictures as the same percept? Various studies have provided insights into how visual information is processed along the "ventral pathway," via both single-cell recordings in monkeys [1, 2] and functional imaging in humans [3, 4]. Interestingly, in humans, the same "concept" of Marilyn Monroe can be evoked with other stimulus modalities, for instance by hearing or reading her name. Brain imaging studies have identified cortical areas selective to voices [5, 6] and visual word forms [7, 8]. However, how visual, text, and sound information can elicit a unique percept is still largely unknown. By using presentations of pictures and of spoken and written names, we show that (1) single neurons in the human medial temporal lobe (MTL) respond selectively to representations of the same individual across different sensory modalities; (2) the degree of multimodal invariance increases along the hierarchical structure within the MTL; and (3) such neuronal representations can be generated within less than a day or two. These results demonstrate that single neurons can encode percepts in an explicit, selective, and invariant manner, even if evoked by different sensory modalities.

Results

We previously reported the presence of single neurons in the human MTL that fire in a highly selective and invariant manner to different pictures of the same familiar person or object [9]. Given the convergence of unimodal and multimodal cortical areas into the MTL [10, 11], we here consider the extent to which single MTL neurons possess an invariant representation of percepts across sensory modalities.

In 16 experimental sessions with 7 subjects implanted with depth electrodes for clinical reasons, we recorded from 750 MTL units (335 single units and 415 multiunits; 46.9 units per session; SD, 21.0; range, 22–93). Of the 750 recorded units, 79 had a significant response to at least one stimulus (see Table S1 available online). For these neurons, we evaluated whether they fired in an invariant manner to the three different pictures of the particular individual or object eliciting a response, and to their written and spoken names (see Experimental Procedures). Two of the responsive units fired exclusively to a text stimulus (and not to pictures or sound) and none of them fired to sound only.

Single-Cell Examples of Multimodal Invariance

Figure 1 shows a neuron in the left anterior hippocampus that fired selectively to three pictures of the television host Oprah Winfrey and to her written (stimulus 56) and spoken (stimulus 73) name. From a nearly silent baseline (mean, 0.06 Hz; SD, 0.21 Hz), it responded with up to 50 Hz almost exclusively to the presentations of Oprah Winfrey, a nearly 1000-fold increase in its firing rate. To a lesser degree, the neuron also fired to the actress Whoopi Goldberg. None of the other responses were significant, including other text and sound presentations.

Figure 2 displays the firing of a neuron in the entorhinal cortex responding selectively to pictures of Saddam Hussein as well as to the text "Saddam Hussein" and his name pronounced by the computer. As in the previous case, this neuron had a relatively low baseline activity (mean, 0.44 Hz; SD, 0.56) and responded with up to 40 Hz to presentations of Hussein. There were no responses to other pictures, texts, or sounds.

Figure 3 shows a neuron in the amygdala selectively activated by photos, text, and sound presentations of one of the researchers performing recordings with the patient at UCLA. The neuron fired with up to 40 Hz from a mean baseline activity of 0.12 Hz (SD, 0.29), a more than a 300-fold increase. Altogether, we found five units responding to one or more researchers performing experiments at UCLA (see Supplemental Data). None of these researchers were previously known to the patient, thus indicating that MTL neurons can form invariant responses and dynamic associations—linking different individuals into the same category "the researchers at UCLA"—within a day or so.

Additional selective responses to pictures, sound, and text presentations are shown in the Supplemental Data.

Population Results

Figure 4A shows the relative number of responsive units, responsive units with visual invariance, and responsive units with responses to sound and text presentations (see also Table S1). There were no significant differences in the relative number of responses among the different MTL areas. Although the degree of visual invariance was not significantly different for the different areas, we found an interesting trend of increasing invariance along the MTL hierarchical structure, with neurons in hippocampus and entorhinal cortex having the largest degree of visual invariance, followed by neurons in amygdala and finally by neurons in parahippocampal cortex. In agreement with this hierarchical structure, responses to picture presentations in the parahippocampal cortex were

*Correspondence: rqqg1@le.ac.uk

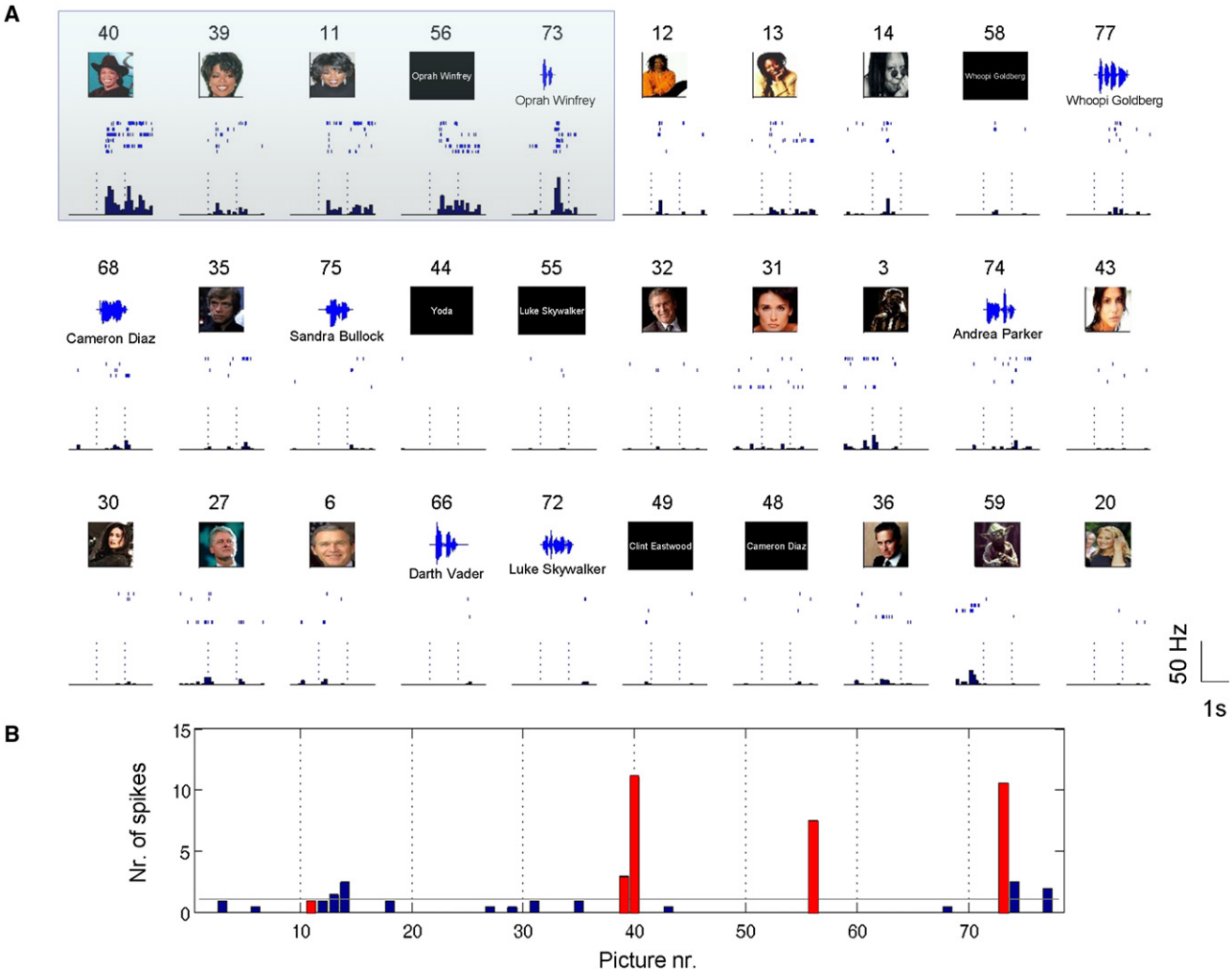

Figure 1. Example of a Neuron with Multimodal Responses to Oprah Winfrey

(A) A neuron in the hippocampus that responded selectively to pictures of the television host Oprah Winfrey (stimulus 40, 39, and 11), as well as to her written (stimulus 56) and spoken (stimulus 73) name. To a lesser degree, the neuron also fired to Whoopi Goldberg. They were no responses to any other picture, sound, or text presentations. For space reasons, only the largest 30 (out of 78) responses are displayed. In each case the raster plots for the six trials, peri-stimulus time histograms (PSTH) and the corresponding pictures are shown. The vertical dotted lines mark picture onset and offset, 1 s apart.

(B) Median number of spikes (across trials) for all stimuli. Presentations of Oprah Winfrey are marked with red bars. Stimulus numbers corresponds to the ones shown above each picture in (A). The gray horizontal line shows the 5 SD above the baseline threshold used for defining significant responses.

significantly earlier than the ones in the other MTL areas (see Supplemental Data and [12]).

The most striking anatomic dissociation occurred for the sound and text responses ($p < 0.01$ and $p < 10^{-4}$, respectively). No neuron in the parahippocampal cortex was activated by any of the sound or text presentations, whereas about a quarter of the responsive neurons in the amygdala and about half of the responsive neurons in the hippocampus and entorhinal cortex responded to sound and text.

Multimodal Invariance

We defined neurons with multimodal "triple invariance" as those having visual invariance together with significant responses to the spoken and written names of the same person or object. Of the 79 responsive units, 17 showed triple invariance according to this definition. In line with the hierarchical organization shown in Figure 4A, between 35% and 40% of the responsive neurons in the hippocampus and entorhinal cortex had multimodal triple invariance. This was the case for only 14% of the neurons in amygdala and for no neuron in the parahippocampal cortex (see Table S1). Ten units (2 in hippocampus, 4 in entorhinal cortex, and 4 in amygdala) had invariant responses involving 2 of the 3 modalities tested: 5 of these responded invariantly to pictures and sound (but not to text), 4 responded invariantly to pictures and text (but not sound), and 1 responded to a picture, the text, and sound presentations of a spider but without visual invariance.

For the 17 units with triple invariance, in Figure 4B we display the average normalized instantaneous firing rate curves for the picture (for each neuron the average over the 3 pictures was used), text, and sound presentations. The instantaneous firing rates were calculated by convolving the spike trains with a Gaussian kernel of 60 ms width and normalizing, for each neuron, to the maximum response (to either picture, text, or

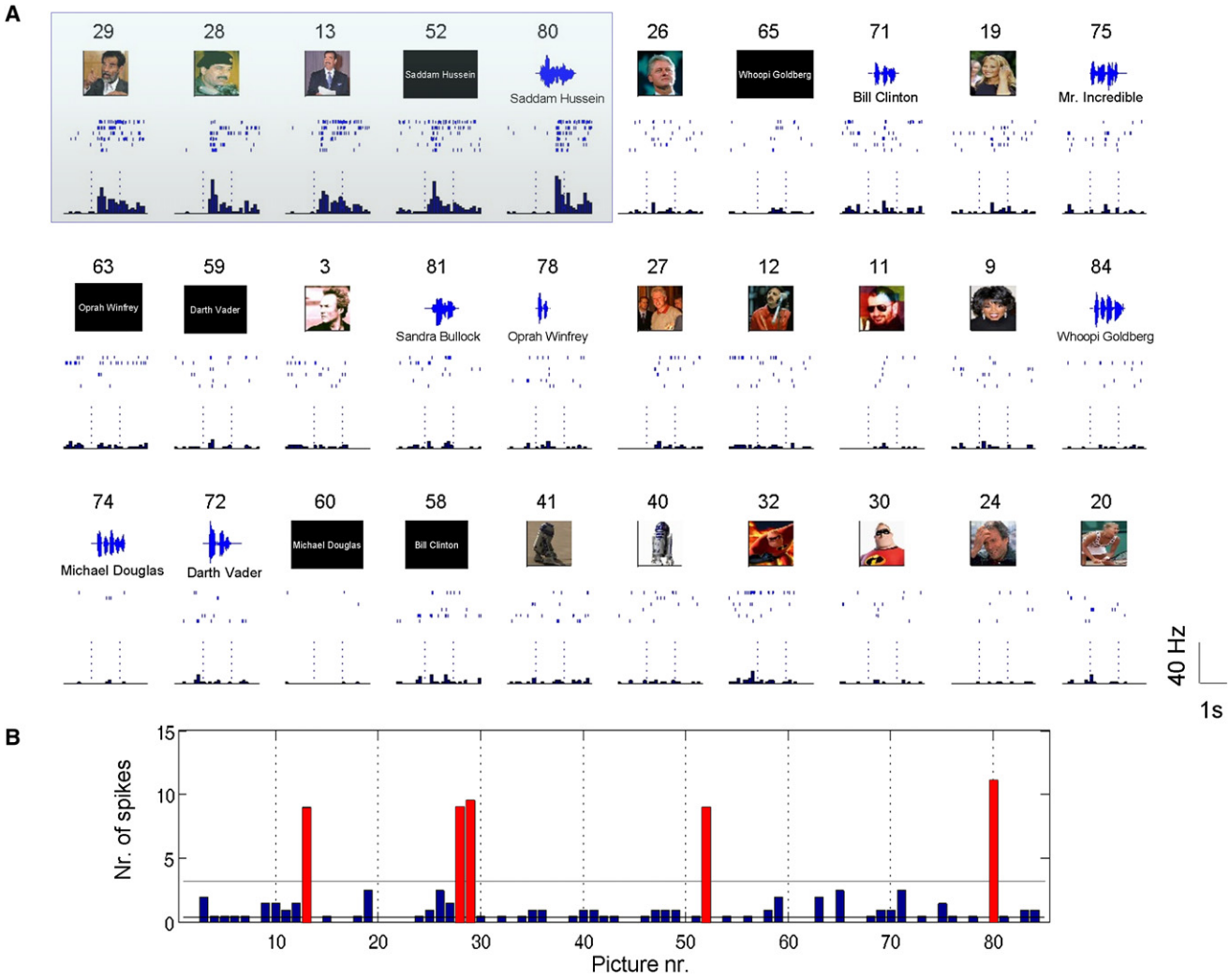

Figure 2. Example of a Neuron with Multimodal Responses to Saddam Hussein

A neuron in the entorhinal cortex that fired selectively to three pictures of Saddam Hussein, his written name, and his spoken name. Conventions are the same as for Figure 1. The black and gray horizontal lines show the mean baseline firing and the threshold for defining significant responses, respectively.

sound). Eleven of the neurons showing triple invariance responded to only one person and the remaining six responded to more than one person or object. To avoid overemphasizing the firing pattern of neurons with more than one response, for each neuron we considered only the person/object eliciting the largest activation. Responses to picture presentations had the earliest responses, followed by those to text and then the ones to sound presentations (see Supplemental Data).

To statistically compare the magnitude of the picture, text, and sound responses, we used the total number of spikes fired in each condition (between 0 and 1000 ms after stimulus onset for the picture and text responses, and between 500 and 1500 ms for the sound responses; see Supplemental Data). A one-way ANOVA test showed that the number of spikes elicited in these three conditions were not significantly different ($p = 0.33$).

Discussion

How the brain integrates information from different sensory modalities has been a topic of extensive research, for example, for the study of orienting behavior—that is, localizing an event perceived by more than one sense (for a recent review see [13])—or the congruence between sight and sound information, as in the "ventriloquist effect" [14]. In particular, single-cell recordings have revealed multisensory neurons in the superior colliculus of cats [15, 16] and in the posterior parietal cortex [17–19], the superior temporal sulcus [20–22], and the prefrontal cortex [23, 24] of monkeys. Neuronal correlates of multisensory processing in cortex have also been reported with EEG and fMRI studies [14, 25–27], but in this case one cannot exclude the possibility that EEG or fMRI responses simply reflect the presence of different populations of unimodal neurons [28]. Our results extend these findings by showing how information from different sensory modalities converges onto neurons in the human MTL. In fact, in the cat (for example), superior colliculus neurons respond to either a visual or an auditory cue signaling a location, and neurons in the human MTL responded to high-level percepts—such as the identity of a given person—triggered by different pictures of the person or by his or her written or spoken name. This convergence of information across different sensory modalities followed the

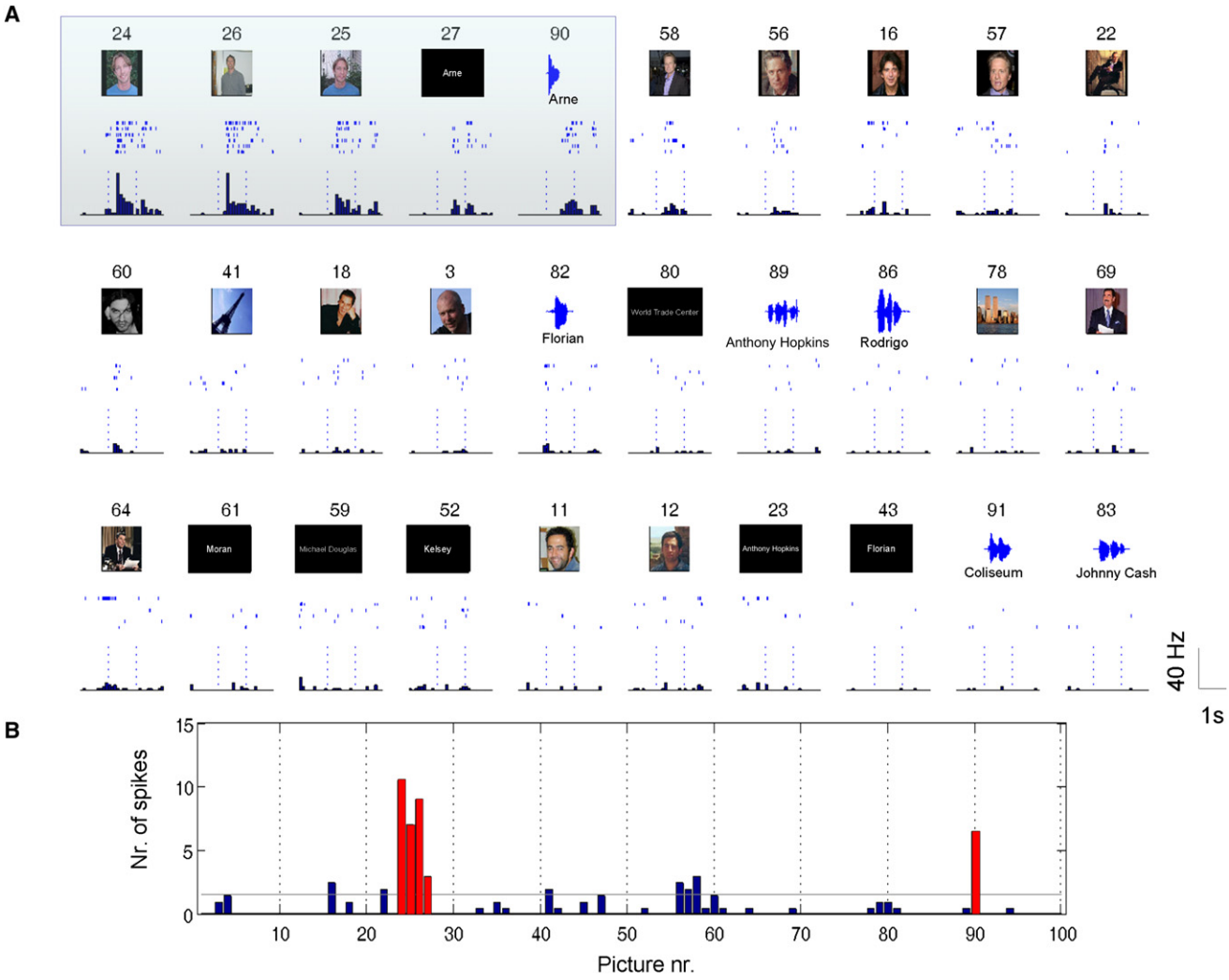

Figure 3. Example of a Neuron with Multimodal Responses to a UCLA Researcher

A neuron in the amygdala selectively activated by pictures, sound, and text presentations of Arne, one of the researchers doing experiments at UCLA. This person was unknown to the patient before the experiments took place. From the same recording electrode, a second neuron firing to pictures of Michael Douglas and a multiunit firing to pictures of a family member of the patient were isolated after spike sorting (see Figures S1 and S6), thus showing that nearby neurons can have very different responses.

hierarchical structure of the MTL, as indicated by the fact that it was not present at the level of parahippocampal cortex and reached its maximum in the entorhinal cortex and hippocampus. In particular, in the parahippocampal cortex, half of the responsive neurons showed visual invariance while *none* fired to the spoken or written names, whereas in hippocampus and entorhinal cortex, three-quarters of the responsive neurons showed visual invariance and half of them responded to sound and text.

Given the previous finding that human MTL neurons can be activated by visual imagery [29], it might be in principle possible that the text and sound activations reported here are just responses to visual imagery. However, this possibility is unlikely for two reasons. (1) It seems more difficult to elicit spontaneous imagery responses with eyes open, as in our experiments, than when subjects are specifically asked to imagine a picture with eyes closed, as in [29]. (2) The responses to the text and sound stimuli were as strong as the ones to the pictures and there was a relatively small delay of less than 100 ms between the responses to the picture and text presentations, likely resulting from the different times required to understand a text and recognize a face. The responses to sound stimuli were much later because in this case the presentation onset is not as clearly defined as with pictures. In contrast to these findings, imagery responses have been reported to be weaker than those to picture presentations and with a more variable and larger delay (the latency of imagery responses was more than 200 ms longer than the one to the picture responses) [29, 30].

Although it is not possible with the current data to provide a conclusive mechanistic explanation of how such abstract single-cell multimodal responses may arise, some insights can be gained by comparing our findings to those described with single-cell recordings in monkeys. Concerning visual invariance, several studies have shown that IT neurons have some—rather limited—degree of invariance, mainly to the size and position of the stimuli [1, 2, 31]. Given the direct inputs from IT to the MTL [10, 11], the distributed representation in IT

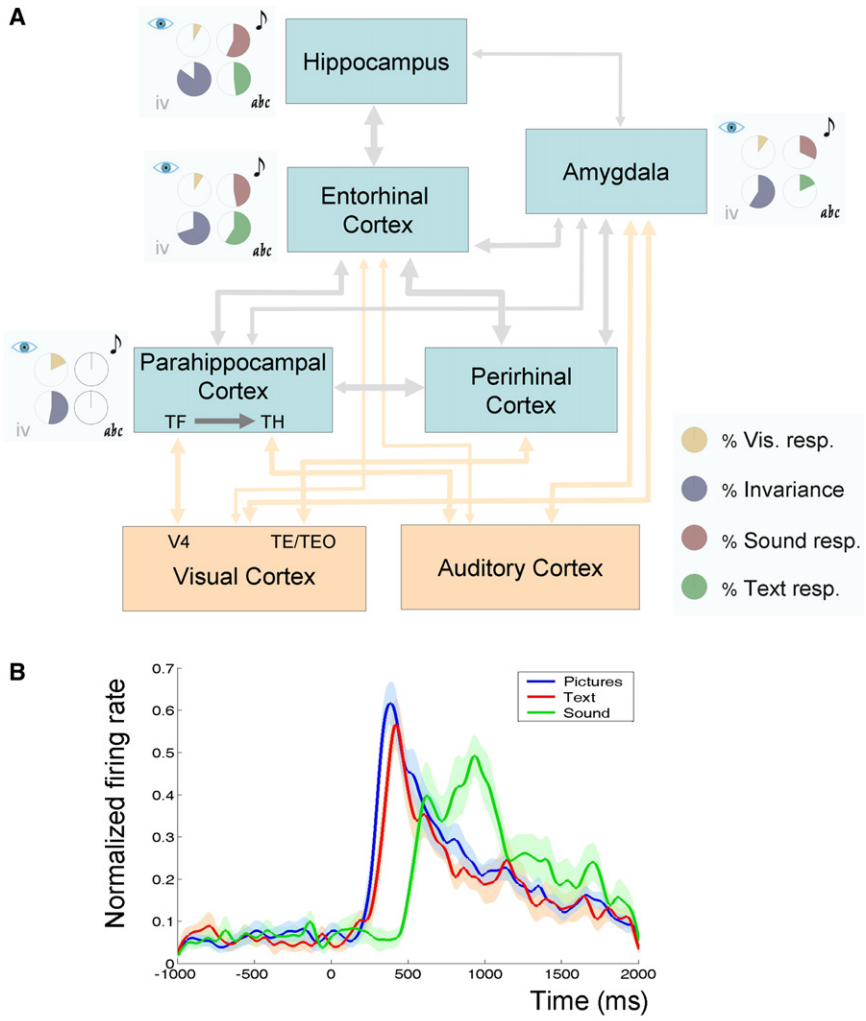

Figure 4. Population Results

(A) Percentage of responsive units, units with visual invariance, and units with responses to sound and text presentations. Anatomical connections between the different MTL areas and with visual and auditory cortices in the macaque monkey [10, 11] are marked with light gray and orange arrows, respectively. Note the increasing degree of visual invariance and number of responses to sound and text presentations along the hierarchical structure of the MTL, with the largest relative number of responses in the hippocampus and entorhinal cortex and the lowest in the parahippocampal cortex.

(B) Normalized firing rates to picture, text, and sound presentations for all 17 neurons with triple invariance (i.e., with responses to picture, sound, and text). Bands around the average values denote SEM.

cortex—in the sense that IT neurons fire to several faces [32]—could generate the much sparser and abstract visual representation of the MTL neurons described here. Extending this to multimodal invariance, it seems plausible that the MTL links information from different sensory modalities given the anatomical convergence of multiple sensory modalities to this area [10, 11, 33]. In this respect, evidence points toward a role of the MTL in forming associations [34, 35] and therefore, it seems reasonable to postulate that the multimodal responses described here can be created through associations, by linking faces with the written and spoken names. Further support to the view that these cells may be encoding associations is given by the fact that some of the units were activated by concepts that were clearly related at an abstract level, such as a neuron firing to two Star Wars characters or a neuron firing to four landmark buildings or a neuron firing to four researchers that performed experiments with the patient (Figures S12, S14, and S8, respectively).

It has been argued that the sparse representation of MTL neurons is particularly suited to allow very fast learning without catastrophic interference [36]. In this respect, it is indeed remarkable that the MTL neurons we recorded from could create invariant responses and associations in less than a day or two, because five of our MTL units (Figure 3; Figures S7–S10) fired selectively to the pictures as well as written and spoken names of researchers performing experiments at UCLA, which were previously unknown to the patients.

Each concept must activate a population of MTL cells, because if out of perhaps a billion MTL cells we do find a neuron firing to Oprah Winfrey (for example), there should be more [9, 37, 38]. Some neurons of this network may have direct visual inputs, some others direct inputs from auditory cortex, and others direct inputs from text-recognition areas. Then, learning the name of a person could be achieved by linking the activity of these neurons, so that a visual, text, or auditory input will activate the whole network encoding the percept. This simple scenario would explain how MTL neurons could quickly encode percepts and respond to different sensory modalities in an explicit, selective, and multimodal invariant manner, as found in our data.

Experimental Procedures

Subjects and Recordings

The data come from 16 sessions in 7 patients with pharmacologically intractable epilepsy, implanted with intracranial electrodes for clinical reasons. Here we report data from sites in the hippocampus, amygdala, entorhinal cortex, and parahippocampal cortex. Each electrode probe had a total of nine microwires at its end, eight active recording channels, and one reference. The differential signal from the microwires was amplified, filtered between 1 and 9000 Hz, and sampled at 28 kHz.

In the recording sessions, an average of 16.7 (SD, 3.1; range, 11–20) individuals or objects were presented. For each of them, three different pictures as well as their names written in the laptop screen and spoken by a computer-synthesized voice were presented in pseudorandom order, six times each (see Supplemental Data).

Data Analysis

From the continuous wide-band data, spike detection and sorting was carried out with Wave_Clus, an adaptive and stochastic algorithm [39] (see Supplemental Data). Significant responses were defined with a heuristic criteria based on the deviation from baseline firing, as in previous studies [9, 40] (see Supplemental Data). For the responsive units, i.e., those firing to at least one stimulus, we evaluated whether they also responded in an invariant manner to three different pictures of the same individual or object

and to the text and sound presentations. Visual invariance, i.e., a preferred firing to the pictures of a given person or object, was quantified with a receiver operator characteristic (ROC) test [9] (see Supplemental Data). Statistical differences in the response characteristics for the different MTL areas were evaluated by a Fisher Exact Test (see Supplemental Data).

Supplemental Data

Supplemental Data include Supplemental Experimental Procedures, 18 figures, and 2 tables and can be found with this article online at http://www.cell.com/current-biology/supplemental/S0960-9822(09)01377-3.

Acknowledgments

We thank E. Behnke, T. Fields, A. Postolova, and K. Laird for technical assistance. This work was supported by grants from NINDS, EPSRC, MRC, the NIMH, DARPA, and the Mathers Foundation.

Received: March 9, 2009
Revised: May 13, 2009
Accepted: June 5, 2009
Published online: July 23, 2009

References

1. Logothetis, N.K., and Sheinberg, D.L. (1996). Visual object recognition. Annu. Rev. Neurosci. *19*, 577–621.
2. Tanaka, K. (1996). Inferotemporal cortex and object vision. Annu. Rev. Neurosci. *19*, 109–139.
3. Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. J. Neurosci. *17*, 4302–4311.
4. Grill-Spector, K., and Malach, R. (2004). The human visual cortex. Annu. Rev. Neurosci. *27*, 649–677.
5. Belin, P., Fecteau, S., and Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. Trends Cogn. Sci. *8*, 129–135.
6. Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., and Pike, B. (2000). Voice-selective areas in human auditory cortex. Nature *403*, 309–312.
7. Cohen, L., Dehaene, S., Naccache, L., Lehericy, S., Dehaene-Lambertz, G., Henaff, M.-A., and Michel, F. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. Brain *123*, 291–307.
8. Vinckier, F., Dehaene, S., Jobert, A., Dubus, J.P., Sigman, M., and Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. Neuron *55*, 143–156.
9. Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. Nature *435*, 1102–1107.
10. Suzuki, W.A. (1996). Neuroanatomy of the monkey entorhinal, perirhinal and parahippocampal cortices: Organization of cortical inputs and interconnections with amygdale and striatum. Semin. Neurosci. *8*, 3–12.
11. Saleem, K.S., and Tanaka, K. (1996). Divergent projections from the anterior inferotemporal area TE to the perirhinal and entorhinal cortices in the macaque monkey. J. Neurosci. *16*, 4757–4775.
12. Mormann, F., Kornblith, S., Quian Quiroga, R., Kraskov, A., Cerf, M., Fried, I., and Koch, C. (2008). Latency and selectivity of single neurons indicate hierarchical processing in the human medial temporal lobe. J. Neurosci. *28*, 8865–8872.
13. Stein, B.E., and Stanford, T.R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. Nat. Rev. Neurosci. *9*, 255–266.
14. Calvert, G.A., Spence, C., and Stein, B.E. (2004). The Handbook of Multisensory Processes (Cambridge, MA: The MIT Press).
15. Stein, B.E., and Meredith, M.A. (1993). The Merging of the Senses (Cambridge, MA: The MIT Press).
16. Meredith, M.A., and Stein, B.E. (1983). Interactions among converging sensory inputs in the superior colliculus. Science *221*, 389–391.
17. Cohen, Y.E., and Andersen, R.A. (2000). Reaches to sounds encoded in an eye-centered reference frame. Neuron *27*, 647–652.
18. Stricanne, B., Andersen, R.A., and Mazzoni, P. (1996). Eye-centered, head-centered, and intermediate coding of remembered sound locations in area LIP. J. Neurophysiol. *76*, 2071–2076.
19. Andersen, R.A. (1997). Multimodal integration for the representation of space in the posterior parietal cortex. Philos. Trans. R. Soc. Lond. B Biol. Sci. *352*, 1421–1428.
20. Benevento, L.A., Fallon, J., Davis, B.J., and Rezak, M. (1977). Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. Exp. Neurol. *57*, 849–872.
21. Barraclough, N.E., Xiao, D., Baker, C.I., Oram, M.W., and Perrett, D.I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. J. Cogn. Neurosci. *17*, 377–391.
22. Bruce, C., Desimone, R., and Gross, C.G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J. Neurophysiol. *46*, 369–384.
23. Sugihara, T., Diltz, M.D., Averbeck, B.B., and Romanski, L.M. (2006). Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. J. Neurosci. *26*, 11138–11147.
24. Fuster, J. (2000). Cross-modal and cross-temporal association in neurons of frontal cortex. Nature *405*, 347–351.
25. Ghazanfar, A.A., and Schroeder, C.E. (2006). Is neocortex essentially multisensory? Trends Cogn. Sci. *10*, 278–285.
26. Macaluso, E., and Driver, J. (2005). Multisensory spatial interactions: A window onto functional integration in the human brain. Trends Neurosci. *28*, 264–271.
27. Lalanne, C., and Lorenceau, J. (2004). Crossmodal integration for perception and action. J. Physiol. (Paris) *98*, 265–279.
28. King, A.J., and Calvert, G.A. (2001). Multisensory integration: Perceptual grouping by eye and ear. Curr. Biol. *11*, R322–R325.
29. Kreiman, G., Koch, C., and Fried, I. (2000). Imagery neurons in the human brain. Nature *408*, 357–361.
30. Kreiman, G. (2002). On the neuronal activity in the human brain during visual recognition, imagery and binocular rivalry. PhD thesis, California Institute of Technology, Pasadena, CA.
31. Tsao, D.Y., and Livingstone, M. (2008). Mechanisms of face perception. Annu. Rev. Neurosci. *31*, 411–437.
32. Kreiman, G., Hung, C.P., Kraskov, A., Quiroga, R.Q., Poggio, T., and DiCarlo, J.J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. Neuron *49*, 433–445.
33. Lavenex, P., and Amaral, D.G. (2000). Hippocampal-neocortical interaction: A hierarchy of associativity. Hippocampus *10*, 420–430.
34. Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature *335*, 817–820.
35. Wirth, S., Yanike, M., Frank, L.M., Smith, A.C., Brown, E.N., and Suzuki, W.A. (2003). Single neurons in the monkey hippocampus and learning of new associations. Science *300*, 1578–1581.
36. Norman, K.A., and O'Reilly, R.C. (2003). Modelling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. Psychol. Rev. *110*, 611–646.
37. Quian Quiroga, R., Kreiman, G., Koch, C., and Fried, I. (2008). Sparse but not 'Grandmother-cell' coding in the medial temporal lobe. Trends Cogn. Sci. *12*, 87–91.
38. Waydo, S., Kraskov, A., Quian Quiroga, R., Fried, I., and Koch, C. (2006). Sparse representation in the human medial temporal lobe. J. Neurosci. *26*, 10232–10234.
39. Quian Quiroga, R., Nadasdy, Z., and Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput. *16*, 1661–1687.
40. Quian Quiroga, R., Reddy, L., Koch, C., and Fried, I. (2007). Decoding visual inputs from multiple neurons in the human temporal lobe. J. Neurophysiol. *98*, 1997–2007.